SIMCOG TECHNOLOGIES

Big Data and Predictive Analytics in Business Use Cases

HAMBURG,
March 10, 2016

SimCog // COMPANY PROFILE

SIMCOG TECHNOLOGIES → SIM + COG

SIMULATION

COGNITION

SimCog Technologies GmbH

- Founded in 2012
- Based in Hamburg
- 6 Employees → Looking for more Data Scientists
- Several Use Cases in the Area of Predictive Analytics
- Mainly Private Equity owned

Difference between Science and Business

- Shorter turnaround-time in business
    - Projects last ~ 3 Month
- Specialized “skill set” of Data Scientists
    - Finding economic relevant problem and external data (Consultant)
    - Explaining methods / results / tests to “uneducated” managers (Sales / Acquisition)
    - Find best statistical method, e.g. Neural Net vs. BDT (Statistician)
    - Efficient programming in small teams (Programmer)
    - Extract relevant data from Databases, APIs, Web crawler, … (Data Analyst)
- Long-Term Jobs and better Payment

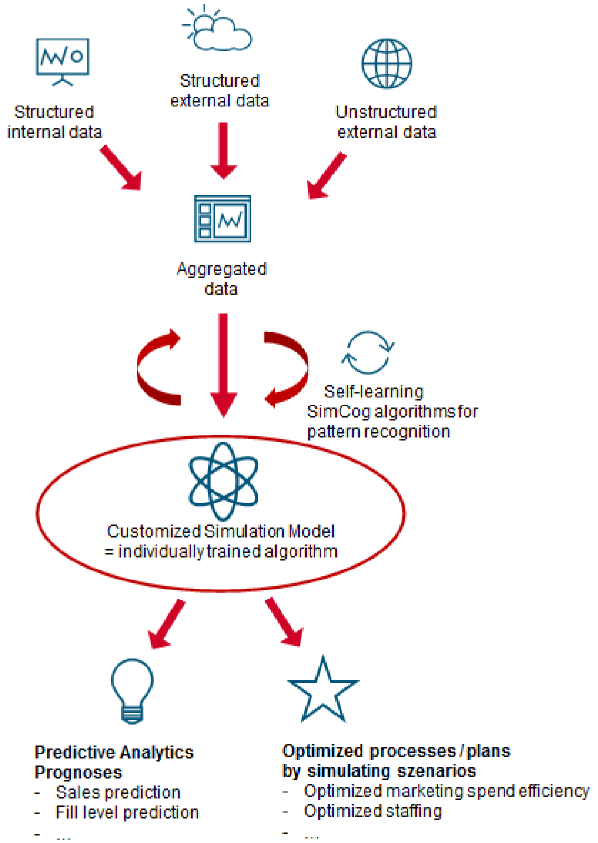
SimCog // ANALYSIS METHOD

Fusion of relevant Data

- Structured internal company data
- Structured external data, e.g. geodata, holiday seasons, weather, etc.
- <u>Special feature:</u> Inclusion of unstructured data is possible, e.g. social media monitoring

Properties of the Simulation Model

- Self-learning algorithms
- Quality checks with historic data
- Selection of appropriate algorithms for specific questions



Structured internal data

Structured external data

Unstructured external data

Aggregated data

Self-learning SimCog algorithms for pattern recognition

Customized Simulation Model = individually trained algorithm

Predictive Analytics Prognoses

- Sales prediction
- Fill level prediction
- ...

Optimized processes / plans by simulating szenarios

- Optimized marketing spend efficiency
- Optimized staffing
- ...

What does “Big Data” mean in Research and Business

Byte	=	Grain of Rice	
Kilobyte	=	Cup of Rice	
Megabyte	=	Eight Bags of Rice	
Gigabyte	**=**	**Three Lorries of Rice**	→ **Business**
Terabyte	=	Two Container ships of Rice	
Petabyte	**=**	**Manhattan covered with Rice**	→ **CERN**
Exabyte	=	Great Britain covered with Rice (3 times)	
Zettabyte	=	Fills Pacific Ocean with Rice	

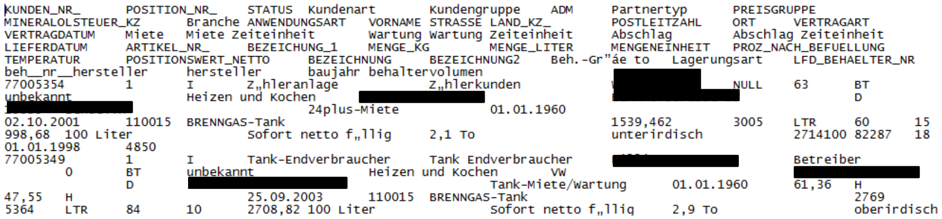
However, data in business comes from many different sources and the inclusion of external data is the key in big data analysis

The Good, the Bad and the Ugly

Some Data-Format Examples

- Most Data available in „good“ CSV-files
  - Inconsistent Data/Time Formats
  - , and . in numbers
  - Missing quotation and binary numers

- Database Dump

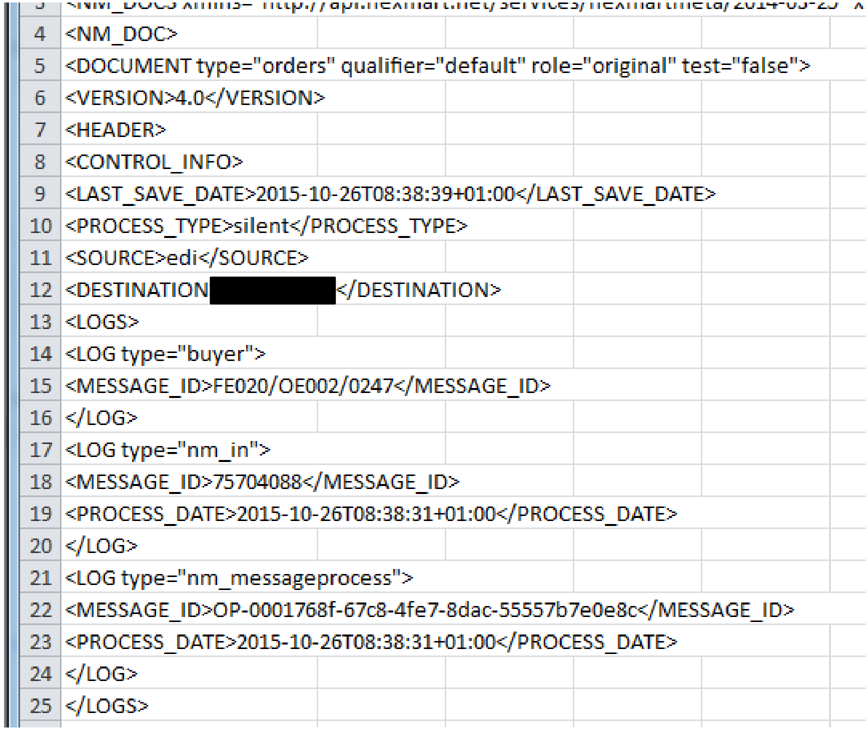


```
KUNDEN_NR_      POSITION_NR_    STATUS  Kundenart       Kundengruppe    ADM     Partnertyp      PREISGRUPPE
MINERALOLSTEUER_KZ      Branche ANWENDUNGSART   VORNAME STRASSE LAND_KZ_        POSTLEITZAHL    ORT     VERTRAGART
VERTRAGDATUM    Miete   Miete Zeiteinheit       Wartung Wartung Zeiteinheit     Abschlag        Abschlag Zeiteinheit
LIEFERDATUM     ARTIKEL_NR_     BEZEICHUNG_1    MENGE_KG        MENGE_LITER     MENGENEINHEIT   PROZ_NACH_BEFUELLUNG
TEMPERATUR      POSITIONSWERT_NETTO     BEZEICHNUNG     BEZEICHNUNG2    Beh.-Gr"áe to   Lagerungsart    LFD_BEHAELTER_NR
beh__nr__hersteller     hersteller      baujahr behaltervolumen
77005354        1       I       Z„hleranlage            Z„hlerkunden            [redacted]      NULL    63      BT
unbekannt               Heizen und Kochen      [redacted]                                                       D
[redacted]                              24plus-Miete            01.01.1960
02.10.2001      110015  BRENNGAS-Tank                                           1539,462        3005    LTR     60      15
998,68  100 Liter               Sofort netto f„llig     2,1 To                  unterirdisch            2714100 82287   18
01.01.1998      4850
77005349        1       I       Tank-Endverbraucher     Tank Endverbraucher     [redacted]              Betreiber
       0        BT      unbekannt               Heizen und Kochen       VW                              [redacted]
                D       [redacted]                              Tank-Miete/Wartung      01.01.1960      61,36   H
47,55   H                       25.09.2003      110015  BRENNGAS-Tank                                           2769
5364    LTR     84      10      2708,82 100 Liter               Sofort netto f„llig     2,9 To                  oberirdisch
```

The Good, the Bad and the Ugly

Some Data-Format Examples

- Most Data available in „good“ CSV-files
  - Inconsistent Data/Time Formats
  - , and . in numbers
  - Missing quotation and binary numers
- Database Dump
- XML in CSV-file

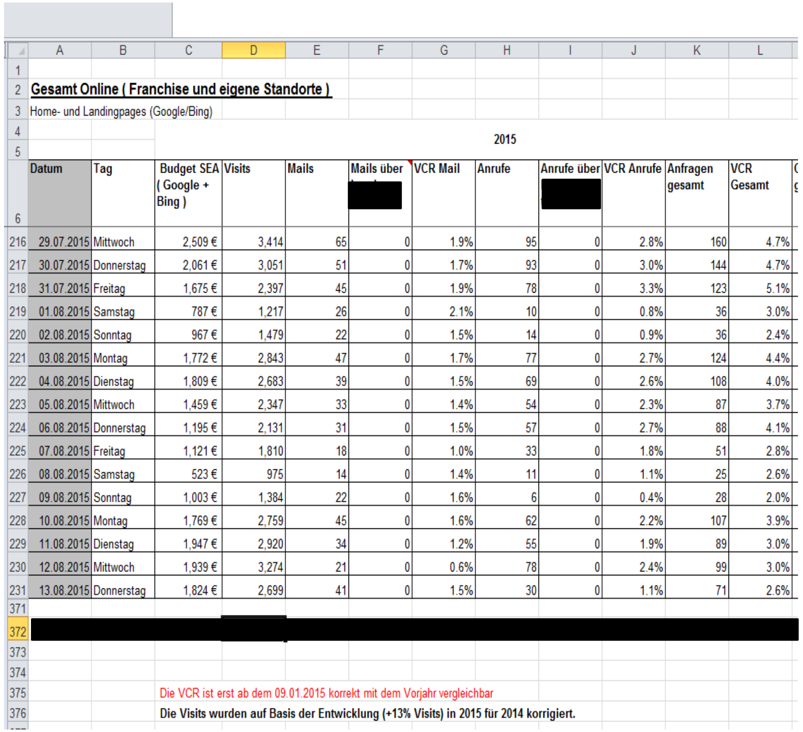


```
<NM_DOCS xmlns="http://api.nexmart.net/services/nexmartmeta/2014-05-23" x
<NM_DOC>
<DOCUMENT type="orders" qualifier="default" role="original" test="false">
<VERSION>4.0</VERSION>
<HEADER>
<CONTROL_INFO>
<LAST_SAVE_DATE>2015-10-26T08:38:39+01:00</LAST_SAVE_DATE>
<PROCESS_TYPE>silent</PROCESS_TYPE>
<SOURCE>edi</SOURCE>
<DESTINATION[redacted]</DESTINATION>
<LOGS>
<LOG type="buyer">
<MESSAGE_ID>FE020/OE002/0247</MESSAGE_ID>
</LOG>
<LOG type="nm_in">
<MESSAGE_ID>75704088</MESSAGE_ID>
<PROCESS_DATE>2015-10-26T08:38:31+01:00</PROCESS_DATE>
</LOG>
<LOG type="nm_messageprocess">
<MESSAGE_ID>OP-0001768f-67c8-4fe7-8dac-55557b7e0e8c</MESSAGE_ID>
<PROCESS_DATE>2015-10-26T08:38:31+01:00</PROCESS_DATE>
</LOG>
</LOGS>
```

The Good, the Bad and the Ugly

Some Data-Format Examples

- Most Data available in „good“ CSV-files
  - Inconsistent Data/Time Formats
  - , and . in numbers
  - Missing quotation and binary numers
- Database Dump
- XML in CSV-file
- Excel Sheets



Gesamt Online (Franchise und eigene Standorte)

Home- und Landingpages (Google/Bing)

2015

	Datum	Tag	Budget SEA (Google + Bing)	Visits	Mails	Mails über [redacted]	VCR Mail	Anrufe	Anrufe über [redacted]	VCR Anrufe	Anfragen gesamt	VCR Gesamt
216	29.07.2015	Mittwoch	2,509 €	3,414	65	0	1.9%	95	0	2.8%	160	4.7%
217	30.07.2015	Donnerstag	2,061 €	3,051	51	0	1.7%	93	0	3.0%	144	4.7%
218	31.07.2015	Freitag	1,675 €	2,397	45	0	1.9%	78	0	3.3%	123	5.1%
219	01.08.2015	Samstag	787 €	1,217	26	0	2.1%	10	0	0.8%	36	3.0%
220	02.08.2015	Sonntag	967 €	1,479	22	0	1.5%	14	0	0.9%	36	2.4%
221	03.08.2015	Montag	1,772 €	2,843	47	0	1.7%	77	0	2.7%	124	4.4%
222	04.08.2015	Dienstag	1,809 €	2,683	39	0	1.5%	69	0	2.6%	108	4.0%
223	05.08.2015	Mittwoch	1,459 €	2,347	33	0	1.4%	54	0	2.3%	87	3.7%
224	06.08.2015	Donnerstag	1,195 €	2,131	31	0	1.5%	57	0	2.7%	88	4.1%
225	07.08.2015	Freitag	1,121 €	1,810	18	0	1.0%	33	0	1.8%	51	2.8%
226	08.08.2015	Samstag	523 €	975	14	0	1.4%	11	0	1.1%	25	2.6%
227	09.08.2015	Sonntag	1,003 €	1,384	22	0	1.6%	6	0	0.4%	28	2.0%
228	10.08.2015	Montag	1,769 €	2,759	45	0	1.6%	62	0	2.2%	107	3.9%
229	11.08.2015	Dienstag	1,947 €	2,920	34	0	1.2%	55	0	1.9%	89	3.0%
230	12.08.2015	Mittwoch	1,939 €	3,274	21	0	0.6%	78	0	2.4%	99	3.0%
231	13.08.2015	Donnerstag	1,824 €	2,699	41	0	1.5%	30	0	1.1%	71	2.6%

[redacted]

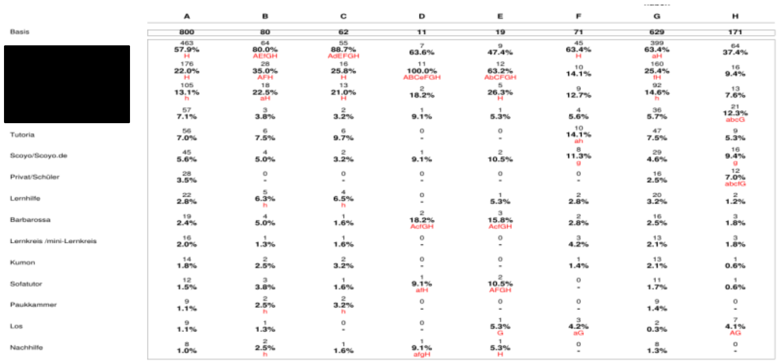
Die VCR ist erst ab dem 09.01.2015 korrekt mit dem Vorjahr vergleichbar

Die Visits wurden auf Basis der Entwicklung (+13% Visits) in 2015 für 2014 korrigiert.

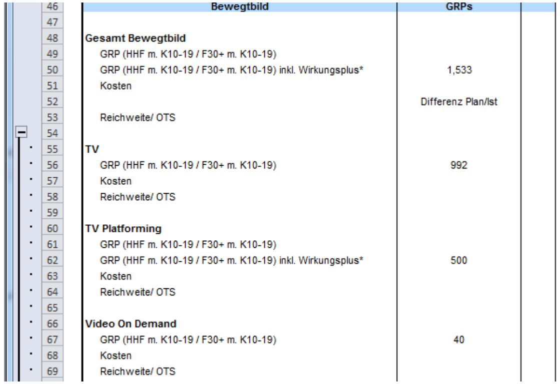
The Good, the Bad and the Ugly

Some Data-Format Examples

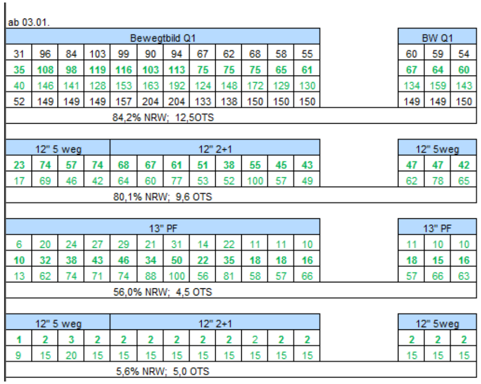
- Most Data available in „good“ CSV-files
  - Inconsistent Data/Time Formats
  - , and . In numbers
  - Missing quotation and binary numers
- Database Dump
- XML in CSV-file
- Excel Sheets
- „Media Plan“-Excel Sheets
- PDF



	A	B	C	D	E	F	G	H
Basis	800	80	62	11	19	71	629	171
[redacted]	463 57.9% H	64 80.0% AEfGH	55 88.7% AdEFGH	7 63.6%	9 47.4%	45 63.4% H	399 63.4% aH	64 37.4%
[redacted]	176 22.0% H	28 35.0% AFH	16 25.8% H	11 100.0% ABCeFGH	12 63.2% AbCFGH	10 14.1%	160 25.4% fH	16 9.4%
[redacted]	105 13.1% h	18 22.5% aH	13 21.0% H	2 18.2%	5 26.3% H	9 12.7%	92 14.6% h	13 7.6%
[redacted]	57 7.1%	3 3.8%	2 3.2%	1 9.1%	1 5.3%	4 5.6%	36 5.7%	21 12.3% abcG
Tutoria	56 7.0%	6 7.5%	6 9.7%	0 -	0 -	10 14.1% ah	47 7.5%	9 5.3%
Scoyo/Scoyo.de	45 5.6%	4 5.0%	2 3.2%	1 9.1%	2 10.5%	8 11.3% g	29 4.6%	16 9.4% g
Privat/Schüler	28 3.5%	0 -	0 -	0 -	0 -	0 -	16 2.5%	12 7.0% abcfG
Lernhilfe	22 2.8%	5 6.3% h	4 6.5% h	0 -	1 5.3%	2 2.8%	20 3.2%	2 1.2%
Barbarossa	19 2.4%	4 5.0%	1 1.6%	2 18.2% AcfGH	3 15.8% AcfGH	2 2.8%	16 2.5%	3 1.8%
Lernkreis /mini-Lernkreis	16 2.0%	1 1.3%	1 1.6%	0 -	0 -	3 4.2%	13 2.1%	3 1.8%
Kumon	14 1.8%	2 2.5%	2 3.2%	0 -	0 -	1 1.4%	13 2.1%	1 0.6%
Sofatutor	12 1.5%	3 3.8%	1 1.6%	1 9.1% afH	2 10.5% AFGH	0 -	11 1.7%	1 0.6%
Paukkammer	9 1.1%	2 2.5% h	2 3.2% h	0 -	0 -	0 -	9 1.4%	0 -
Los	9 1.1%	1 1.3%	0 -	0 -	1 5.3% G	3 4.2% aG	2 0.3%	7 4.1% AG
Nachhilfe	8 1.0%	2 2.5% h	1 1.6%	1 9.1% afgH	1 5.3% H	0 -	8 1.3%	0 -



	Bewegtbild	GRPs	
46			
47			
48	**Gesamt Bewegtbild**		
49	GRP (HHF m. K10-19 / F30+ m. K10-19)		
50	GRP (HHF m. K10-19 / F30+ m. K10-19) inkl. Wirkungsplus*	1,533	
51	Kosten		
52		Differenz Plan/Ist	
53	Reichweite/ OTS		
54			
55	**TV**		
56	GRP (HHF m. K10-19 / F30+ m. K10-19)	992	
57	Kosten		
58	Reichweite/ OTS		
59			
60	**TV Platforming**		
61	GRP (HHF m. K10-19 / F30+ m. K10-19)		
62	GRP (HHF m. K10-19 / F30+ m. K10-19) inkl. Wirkungsplus*	500	
63	Kosten		
64	Reichweite/ OTS		
65			
66	**Video On Demand**		
67	GRP (HHF m. K10-19 / F30+ m. K10-19)	40	
68	Kosten		
69	Reichweite/ OTS		



ab 03.01.

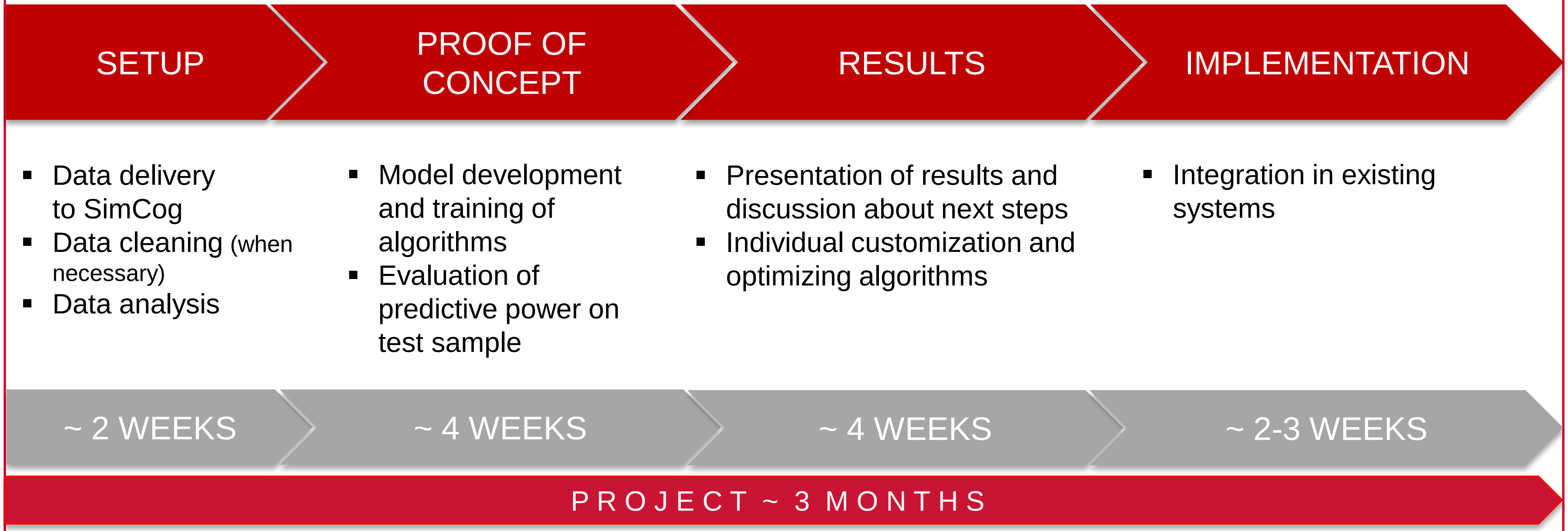
Bewegtbild Q1												BW Q1		
31	96	84	103	99	90	94	67	62	68	58	55	60	59	54
35	108	98	119	116	103	113	75	75	75	65	61	67	64	60
40	146	141	128	153	163	192	124	148	172	129	130	134	159	143
52	149	149	149	157	204	204	133	138	150	150	150	149	149	150
84,2% NRW; 12,5OTS														
12" 5 weg				12" 2+1								12" 5weg		
23	74	57	74	68	67	61	51	38	55	45	43	47	47	42
17	69	46	42	64	60	77	53	52	100	57	49	62	78	65
80,1% NRW; 9,6 OTS														
13" PF												13" PF		
6	20	24	27	29	21	31	14	22	11	11	10	11	10	10
10	32	38	43	46	34	50	22	35	18	18	16	18	15	16
13	62	74	71	74	88	100	56	81	58	57	66	57	66	63
56,0% NRW; 4,5 OTS														
12" 5 weg				12" 2+1								12" 5weg		
1	2	3	2	2	2	2	2	2	2	2	2	2	2	2
9	15	20	15	15	15	15	15	15	15	15	15	15	15	15
5,6% NRW; 5,0 OTS														

Programming Languages

- C++
- ROOT (only at SimCog)
- Java (Natural Language Processing, APIs)
- R
- Python
- Database (SQL, noSQL etc.)

First Steps before Starting a Project

- Task: Find an economic relevant problem, that can be answered with the company data*
- Required prediction quality depends on economic leverage



SETUP	PROOF OF CONCEPT	RESULTS	IMPLEMENTATION
▪ Data delivery to SimCog ▪ Data cleaning (when necessary) ▪ Data analysis	▪ Model development and training of algorithms ▪ Evaluation of predictive power on test sample	▪ Presentation of results and discussion about next steps ▪ Individual customization and optimizing algorithms	▪ Integration in existing systems
~ 2 WEEKS	~ 4 WEEKS	~ 4 WEEKS	~ 2-3 WEEKS

PROJECT ~ 3 MONTHS

*Typically companies have historical data of the past ~ 4 years

SimCog // PREDICTIVE ANALYTICS FILL LEVEL PREDICTION

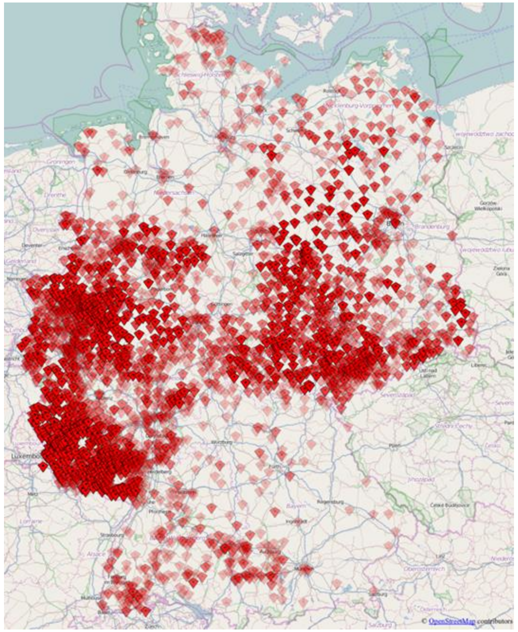
USE CASE ENERGY

Initial question

- Client is an energy supplier that delivers gas to ~ 30.000 consumption points in Germany
- Basic idea: optimize delivery logistics by a precise fill level prediction

Used internal data → *provided by client*

- All gas deliveries since 2002
  - Quantity delivered, date of delivery, fill level after filling, …
- Contract type
  - Heating, heating & cooking, balloonist, industry, …
- Fuel tank capacity
- Use of alternative energy sources
- Postal code



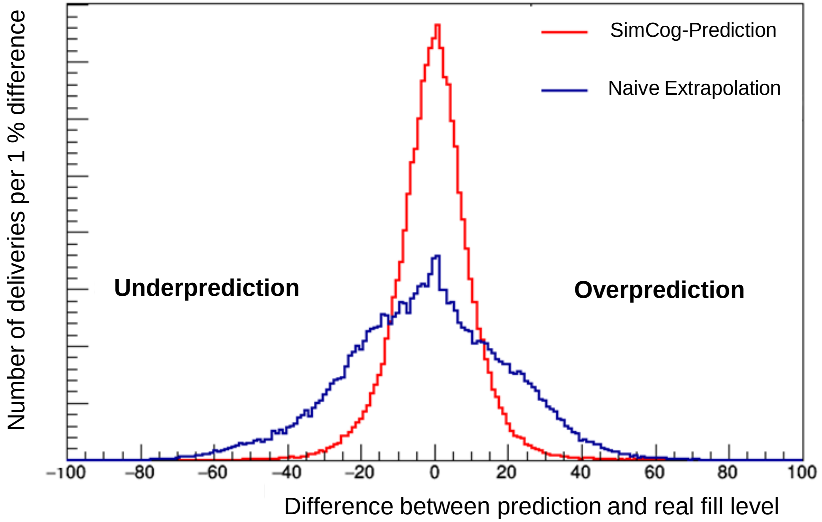
Used external data → *added by SimCog*

- Weather (per consumption point weather information from the three nearest weather stations is used)
- Geo-structural data, e.g. population density
- Information of the tank supplier

SimCog // PREDICTION RESULTS AND QUALITY

USE CASE ENERGY

Comparison Naive Extrapolation / SimCog-Prediction



Comparison of predictions:

The method „Naive Extrapolation“ considers the past and carries on the average consumption of the past

The SimCog-Prediction uses latest machine-learning-algorithms and integrates external data

→ SimCog´s pattern recognition is more than twice as accurate as the naive extrapolation

→ **Precision**
Very precise predictions: uncertainty ~ 5% on sales volumes

SimCog // PREDICTIONS - OUTPUT

USE CASE ENERGY

- Fill level prediction
  - Daily updates of the fill level prediction for 30’000 existing customers
- Delivery date prediction
  - Daily updates of the delivery date prediction for 30’000 existing customers
- Identification of illegal third-party refillment
- Quarterly prediction for optimized supply chain management, energy disposition, purchasing & liquidity planning (of the client)
- Detection of essential data errors
  - Issue warnings for unexpected events such as: negative consumption, illegal third-party refillment, strong behavior change, data error, …
- Automatic adjustment of predictions when behavior is altered (self-learning)

SimCog // BENEFITS

USE CASE ENERGY

1. Reduced storage and standby costs
   - Optimized supply chain management through greatly improved estimated amount of energy needed
2. Cost savings through improved route planning
   - Optimized distribution logistics
   - Better utilization of refilling trucks
   - Customers can be addressed directly for e.g. „early refueling“
3. Improved marketing and sales activities
   - Improved timing of customer contact
   - Special promotion more controllable
4. Improved handling of illegal third-party refillment
   - Phone customers where the probablility of third-party refillment is high
   - Legal action is an option in safe cases
5. Improved overview of turnover by predicting the fill level in the counter systems

SimCog // USE CASES

Retail

Retail B2C

- Sales Prediction
- Footfall
- Customer Forecast
- Shopping Cart
- Online / Offline

Retail B2B

- Sales Prediction
- Online / Offline

Marketing

Marketing

- Marketing Spend Efficiency
- Retargeting
- Churn Rate
- Coupon-Conversion

Logistics

Logistics

- Shipping ETA
- Fill Level Prediction Energy

More Solutions

More Solutions

- Fraud Detection
- Stock Price Prediction
- Protection against Economic Espionage
- …

SimCog // STOCK PRICE PREDICTION

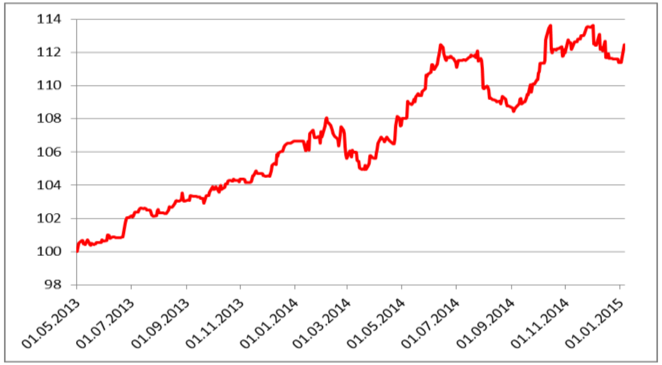
USE CASE STOCK PRICE

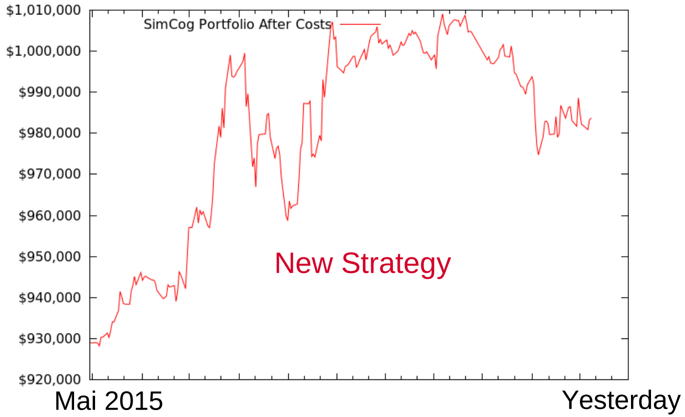
Predictive Analytics: Stock Price

Use Case

- Predict most likely stock price movements and trade with an automated market-neutral trading strategy
- In addition to using stock prices, sector information, director dealings, etc. the analysis of social media information is a crucial factor
  - Discussion from the social web are considered within individual subject areas

Result of real trades on SimCog account:





SimCog // CONTACT

Dr. Jan Thomsen
Managing Director

SimCog Technologies GmbH
Schauenburgerstr. 27
20095 Hamburg

T: +49 40 2000 396 07
F: +49 40 2000 396 20
E: thomsen@simcog.de

www.simcog.de